The $M/M/1$ queue in a heavy-tailed random environment

O.J. Boxma[1,2,*] I.A. Kurkova[1]

1. EURANDOM
P.O. Box 513, 5600 MB Eindhoven
The Netherlands

2. Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven
The Netherlands

Abstract

We consider an $M/M/1$ queue with the special feature that the speed of the server alternates between two constant values s_L and $s_H > s_L$. The high-speed periods are exponentially distributed, and the low-speed periods have a regularly varying distribution. We obtain explicit asymptotics for the tail of the workload distribution. The two cases in which the offered traffic load is smaller respectively larger than the low service speed are shown to result in completely different asymptotics.

1 Introduction and model description

The most extensively studied queueing system is the $M/M/1$ queue: The single server queue fed by a Poisson arrival process, customers requiring exponentially distributed service times. In this paper we study this $M/M/1$

[*] also: CWI, P.O.Box 94079, 1090 GB Amsterdam, The Netherlands, boxma@win.tue.nl; kurkova@eurandom.tue.nl

queue, with the additional feature that the speed of the server is not constant; it alternates between a low speed s_L and a high speed $s_H > s_L$. Our main object of study is the steady-state distribution of the workload V, the amount of work present in the system. In particular, we are interested in the behaviour of $P(V > v)$ for large v. Some reflection shows that one must distinguish between two different regimes in studying these workload asymptotics. Introducing the offered load $\rho := \lambda/\mu$, where λ denotes the arrival rate and $1/\mu$ the mean service request, these regimes can be described as follows.

(i) $\rho > s_L$: If the server always operates at low speed, the workload grows indefinitely.

(ii) $\rho < s_L$: The server can handle all offered work even if it always operates at low speed.

We refrain in this paper from discussing the more delicate case $\rho = s_L$. We shall assume that the periods of low and high speed are independent of each other and of the traffic processes, that the high-speed periods are exponentially distributed, and that the low-speed period distributions have a non-exponential, regularly varying tail (see Bingham, Goldie and Teugels (1987)). Roughly speaking,

$$P(L > t) \sim Ct^{-\zeta}, \qquad t \to \infty,$$

where $f(t) \sim g(t)$ for $t \to \infty$ means that $\lim_{t\to\infty} f(t)/g(t) = 1$.

In an ordinary fixed-speed $M/M/1$ queue with speed s_L, one has, cf. Cohen (1982):

$$P(V > v) = \frac{\lambda}{s_L \mu} \mathrm{e}^{(\lambda/s_L - \mu)v}, \qquad v > 0.$$

In the present model, it will turn out that

(i) if $\rho < s_L$, then $P(V > v) \sim C_1 v^{1-\zeta} \mathrm{e}^{(\lambda/s_L - \mu)v}, \quad v \to \infty;$

(ii) if $\rho > s_L$, then $P(V > v) \sim C_2 v^{1-\zeta}, \quad v \to \infty.$

The main goal of the paper is to prove and explain these results, which expose a completely different workload tail behaviour in the two traffic regimes. In fact, the results to be obtained are slightly more detailed; we also consider the tail behaviour during low-speed and during high-speed periods, respectively.

Motivation of the study. Our motivation for this study is two-fold. Firstly, the single-server queue with various speeds is a very important model, which arises naturally in, e.g., the performance analysis of integrated-services communication networks. In such networks, the influence of high-priority traffic on low-priority traffic is often reflected in a variable capacity (speed) for low-priority traffic. Examples are provided by ABR (Available Bit Rate) traffic and by scheduling disciplines like GPS (Generalized Processor Sharing). As a design paradigm, GPS is at the heart of commonly-used scheduling algorithms for high-speed switches such as Weighted Fair Queueing. From a queueing point of view, GPS gives rise to the analysis of coupled servers, where the speed of one server depends on whether another server is busy or idle. Concentrating on one server, one then sees an alternation of two different speeds. The asymptotic behaviour of the coupled processors was studied by Borst, Boxma and Jelenković (1999). The present model is in one sense more general: The low-speed periods do not necessarily correspond to busy periods of another queue.

A second motivation for this study is the convincing evidence of long-tailed traffic characteristics in high-speed communication networks. Early indications of the long-range dependence of Ethernet traffic, attributed to long-tailed file size distributions, were reported in Leland, Taqqu, Willinger and Wilson (1993). Long-tailed characteristics of the scene length distribution of MPEG video streams were investigated in Heyman and Lakshman (1996) and Jelenković, Lazar and Sermet (1997). These and other empirical findings have triggered theoretical developments in the modeling and queueing analysis of long-tailed traffic phenomena. The influence of long-tailed service time distributions on waiting time and workload distributions of the single-server queue has been investigated in considerable detail; many results are gathered in the book edited by Park and Willinger (2000).

Related work. There is a considerable literature on the single-server queue with several service speeds. An early paper is due to Yechiali and Naor (1971). They have studied the $M/M/1$ queue which alternates between two (see Yechiali (1973) for an extension) exponentially distributed phases, the arrival and service rates depending on the phase. Neuts (1971) has generalized their study to the $M/G/1$ case. He deviates from the assumptions in

[24] by assuming that the service time distribution of a customer depends only on the state of the phase process *at the time his service begins*. Halfin (1972) analyzes the buffer content of an $M/G/1$ queue whose service rate varies according to a birth-and-death process with $c+1$ states. A system of Volterra-type integral equations is derived for the joint distribution of the buffer content and the phase of the birth-and-death process, and is used for the numerical calculation of the distribution. Several authors have considered queues with service interruptions. In our setting this corresponds to taking $s_L = 0$. Some recent studies concerning such queues are Takine and Sengupta (1997), Li, Shi and Chao (1997) and Núñez Queija (1998).

The present paper builds upon the previous work of Boxma and Kurkova (1999). There the $M/G/1$ case has been studied, with regularly varying service time distribution (which does not contain the case of an exponential service time distribution). An interesting feature of the present study is the interplay between the large deviations behaviour of an $M/M/1$ queue and the large deviations behaviour of the low-speed periods. A similar interplay was discussed in Boxma, Deng and Zwart (1999), where the workload was studied in an $M/G/2$ queue where the service time distribution at one server is exponential and at the other server regularly varying.

Set-up of the paper. The paper is organized as follows. The main result is presented in Theorem 1. It gives the tail asymptotics of the workload distribution during low-speed and high-speed periods, both for the case $\rho < s_L$ and the case $\rho > s_L$. Corollary 1 subsequently gives the overall workload asymptotics, i.e., the tail behaviour of the workload distribution at an arbitrary epoch. The remainder of Section 2 is devoted to a discussion of the asymptotic results; intuitive explanations of the results are provided. Section 3 contains the proof of Theorem 1. We end the present section with a detailed model description.

Model description. We consider the $M/M/1$ queue with an infinite buffer. Customers arrive according to a Poisson process with rate λ. The required service times have an exponential distribution with mean $1/\mu$. The speed of the server alternates between two constant values s_L and $s_H > s_L$. The high-speed periods are exponentially distributed with mean $1/\nu$. The low-speed periods have distribution $L(t)$ with Laplace-Stieltjes transform (LST) $\delta(s)$, Re $s \geq 0$, and mean δ. All interarrival times, service requests, lengths of high-speed periods and lengths of low-speed periods are independent. We also assume that the necessary and sufficient condition for stability of this system, which is

$$\frac{\lambda}{\mu} < \frac{\delta}{\delta + 1/\nu} s_L + \frac{1/\nu}{\delta + 1/\nu} s_H,$$

holds.

A key feature of the model is that the low-speed periods have a heavy-tailed distribution, as specified in the following assumption.

Assumption 1. We assume that $\delta(s)$ can be represented in $\mathrm{Re}\, s \geq 0$ as:

$$\frac{1 - \delta(s)}{s} = \delta + s^{\zeta - 1}[C_L \Gamma(1 - \zeta) + l(s)] + h(s), \tag{1}$$

where

(i) $1 < \zeta < 2$;

(ii) $h(s)$ is analytic in $\{s : \mathrm{Re}\, s > -\varepsilon_0\}$ for some $\varepsilon_0 > 0$, $h(0) = 0$;

(iii) $l(s)$ is analytic in $\{s : \mathrm{Re}\, s > 0 \text{ or } |s| < \varepsilon_0\}$ for some $\varepsilon_0 > 0$ and continuous in $\{s : \mathrm{Re}\, s \geq 0\}$, $l(0) = 0$.

According to Theorem 8.1.6 of Bingham, Goldie and Teugels (1987), Assumption 1 implies that $1 - L(t)$ varies regularly with index $-\zeta$. Namely,

$$1 - L(t) = t^{-\zeta}[C_L + o(1)] l_0(t) \qquad \text{as } t \to \infty, \tag{2}$$

where $l_0(t)$ is a slowly varying function at infinity, cf. Bingham, Goldie and Teugels (1987). Various examples of distributions satisfying Assumption 1 are presented in Boxma and Cohen (1999).

Remark 1. The workload distribution in an $M/M/1$ queue with arrival rate λ, service rate μ and speed s_L is readily seen to coincide with the workload distribution in an $M/M/1$ queue with arrival rate $\lambda s_L / s_H$, service rate μ and speed s_H. Hence our model with two service speeds can be translated into a model with one service speed but two arrival rates. In order to find the workload distribution in the model with two service speeds from the workload distribution in the model with two arrival rates, one has to take appropriate weight factors for the two different arrival rate periods.

2 Buffer content distribution

In this section we consider the tail behaviour of the workload (or buffer content) distribution. We do not attempt to obtain the exact workload distribution. That is a complicated problem of its own; see Boxma and Kurkova (1999) for a discussion of its complexity and for an exact analysis for a particular class of low-speed distributions. Let V be the buffer content (workload) in the stationary regime. Let $X = H$ (respectively $X = L$) whenever the speed of the service is s_H (s_L). Denote by L_{past} the time that passed since the last change of service speeds from s_H to s_L. Let us introduce the distribution functions of the workload at high-speed and low-speed periods:

$$\begin{aligned} F_H(v) &:= P(V \leq v,\ X = H), \\ F_L(v,\eta)\mathrm{d}\eta &:= P(V \leq v,\ X = L,\ \eta < L_{\mathrm{past}} \leq \eta + \mathrm{d}\eta), \\ F_L(v) &:= P(V \leq v, X = L). \end{aligned}$$

Theorem 1 (i) *Let $\lambda/\mu > s_L$. If Assumption 1 holds, then*

$$\frac{1}{1+\nu\delta} - F_H(v) \sim D_H v^{1-\zeta}, \tag{3}$$

$$\frac{\nu(1-L(\eta))}{1+\nu\delta} - F_L(v,\eta) \sim \nu(1-L(\eta))D_H v^{1-\zeta}, \tag{4}$$

$$\frac{\nu\delta}{1+\nu\delta} - \int\limits_{\eta=0}^{\infty} F_L(v,\eta)\mathrm{d}\eta \sim D_L v^{1-\zeta}, \tag{5}$$

as $v \to \infty$, where

$$\begin{aligned} D_H &= \frac{\nu(\lambda/\mu - s_L)^{\zeta} C_L}{(1+\nu\delta)((s_H - \lambda/\mu) + \delta\nu(s_L - \lambda/\mu))}\frac{1}{\zeta - 1}, \\ D_L &= \frac{\nu(\lambda/\mu - s_L)^{\zeta-1}(s_H - \lambda/\mu) C_L}{(1+\nu\delta)((s_H - \lambda/\mu) + \delta\nu(s_L - \lambda/\mu))}\frac{1}{\zeta - 1}. \end{aligned}$$

(ii) *Let $\lambda/\mu < s_L$. If Assumption 1 holds, then*

$$\frac{1}{1+\nu\delta} - F_H(v) \sim D_H v^{-\zeta}\exp\{(\lambda/s_L - \mu)v\}, \tag{6}$$

$$\frac{\nu(1-L(\eta))}{1+\nu\delta} - F_L(v,\eta) \sim \nu(1-L(\eta))D_H v^{-\zeta}\exp\{(\lambda/s_L - \mu)v\}, \tag{7}$$

$$\frac{\nu\delta}{1+\nu\delta} - \int_{\eta=0}^{\infty} F_L(v,\eta)\mathrm{d}\eta \quad \sim \quad D_L v^{1-\zeta}\exp\{(\lambda/s_L - \mu)v\}, \tag{8}$$

as $v \to \infty$, where

$$\begin{aligned} D_H &= \frac{\nu s_L C_L}{\mu(1+\nu\delta)(s_H - s_L)}\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}, \\ D_L &= \frac{\nu\lambda C_L}{\mu s_L(1+\nu\delta)(\zeta-1)}\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}. \end{aligned}$$

Corollary 1 (i) *Let $\lambda/\mu > s_L$. If Assumption 1 holds, then*

$$P(V > v) \sim \frac{\nu(s_H - s_L)(\lambda/\mu - s_L)^{\zeta-1}C_L}{(1+\nu\delta)((s_H - \lambda/\mu) + \nu\delta(s_L - \lambda/\mu))}v^{1-\zeta}. \tag{9}$$

(ii) *Let $\lambda/\mu < s_L$. If Assumption 1 holds, then*

$$P(V > v) \sim \frac{\nu\lambda C_L\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}}{\mu s_L(1+\nu\delta)(\zeta-1)}v^{1-\zeta}\exp\{(\lambda/s_L - \mu)v\}. \tag{10}$$

Remark 2. Simultaneously with the present study, Borst and Zwart [5] have analysed the workload tail behaviour for a broad class of queues with a mixture of light-tailed and heavy-tailed input flows. They consider both instantaneous and fluid input. Their results contain (10).

Discussion of the results for $\lambda/\mu > s_L$. For large values of v, small workload jumps caused by arrivals are hardly 'visible', and the global picture of the workload behaviour is that of a *fluid* queue, fed by a single on/off source. When that source is on, the workload increases linearly at rate $\lambda/\mu - s_L$ (this corresponds to a positive drift during the low-speed periods). When the source is off, the workload decreases linearly at rate $s_H - \lambda/\mu$ (this corresponds to a negative drift during the high-speed periods). As in the low/high-speed $M/M/1$ queue, the off-periods are exp(ν) distributed, while the on-periods have distribution $L(\cdot)$. The paper of Kella and Whitt (1992) discusses this fluid queue. It points out that (due to PASTA), the workload at off-periods is distributed like the waiting time in an $M/G/1$ queue with arrival rate ν, service speed $s_H - \lambda/\mu$ and service requests $(\lambda/\mu - s_L)L_i$, where the L_i have distribution $L(\cdot)$; these service requests represent the workload increments during on-periods. It is well-known for an $M/G/1$ queue with arrival rate ν, required service time distribution $B(\cdot)$ with mean β and service rate c, cf. Cohen (1973), that the tail of its steady-state workload distribution $W(\cdot)$ has the following asymptotics (B_{res} denotes a residual required service time):

$$1 - W(v) \sim \frac{\nu\beta}{c-\nu\beta} P(B_{\mathrm{res}} > v) = \frac{\nu\beta}{c-\nu\beta} \int_v^\infty \frac{1-B(u)}{\beta}\,\mathrm{d}u. \tag{11}$$

Since the tail of $L(\cdot)$ is regularly varying with index $-\zeta$, the workload of the $M/G/1$ queue described above has the following tail asymptotics:

$$\begin{aligned} 1 - W_{\mathrm{M/G/1}}(v) &\sim \frac{\nu(\lambda/\mu - s_L)\delta}{s_H - \lambda/\mu - \nu\delta(\lambda/\mu - s_L)} P\left(L_{\mathrm{res}} > v(\lambda/\mu - s_L)^{-1}\right) \\ &\sim \frac{\nu(\lambda/\mu - s_L)(\lambda/\mu - s_L)^{\zeta-1} C_L}{(s_H - \lambda/\mu - \nu\delta(\lambda/\mu - s_L))(\zeta - 1)} v^{1-\zeta} \\ &= D_H(1+\nu\delta)v^{1-\zeta} \sim 1 - \frac{F_H(v)}{F_H(\infty)}. \end{aligned}$$

Thus one can see that the tail asymptotics of the workload at high-speed periods in our two-speed system coincide with the tail asymptotics of the workload in the $M/G/1$ system that was just introduced, or equivalently, with the tail asymptotics of the workload in the off-periods of the fluid queue. Kella and Whitt (1992) also observe that the workload of the fluid queue at on-periods is distributed like the sum of the above $M/G/1$ waiting time and the residual part of the workload increment during an on-period, to be denoted by $(\lambda/\mu - s_L)L_{\mathrm{res}}$. It is readily seen that the tail behaviour of this sum is given by

$$\begin{aligned} & 1 - W_{\mathrm{M/G/1}}(v) + P((\lambda/\mu - s_L)L_{\mathrm{res}} > v) \\ &\sim D_H(1+\nu\delta)v^{1-\zeta} + \frac{C_L(\lambda/\mu - s_L)^{\zeta-1} v^{1-\zeta}}{\delta(\zeta-1)} \\ &= \frac{C_L(\lambda/\mu - s_L)^{\zeta-1}(s_H - \lambda/\mu)}{\delta((s_H - \lambda/\mu) - \nu\delta(\lambda/\mu - s_L))(\zeta - 1)} \\ &= \frac{D_L(1+\nu\delta)v^{1-\zeta}}{\nu\delta}. \end{aligned}$$

This is exactly the same tail behaviour as that of $1 - F_L(v)/F_L(\infty)$ as given in (5). Finally, weighing the workload distributions of the fluid queue during on- and off-periods with weight factors $\nu\delta/(1+\nu\delta)$ and $1/(1+\nu\delta)$, one obtains the overall tail asymptotics of the fluid queue (which have also been derived in Jelenković and Lazar (1999), see also Agrawal, Makowski and Nain (1999)), which of course now also agree with the overall tail asymptotics of the workload in our $M/M/1$ queue with two speeds (cf. (9)). Concluding, the tail asymptotics of the workload in this $M/M/1$ queue and in the fluid queue are exactly the same.

Remark 3. A more global interpretation of (5) is that the most likely way to reach a very high workload is to be in the middle of a very long low-speed (overflow) period; i.e., to have had a very long L_{past}. Similarly, the $v^{1-\zeta}$ behaviour during a high-speed period (cf. (3)) is obtained by considering a low-speed period in the past that has been so long that the present workload still exceeds v (an argument like this has been worked out in detail in Section 4.1 of Boxma, Deng and Zwart (1999)).

Discussion of the results for $\lambda/\mu < s_L$. We can rewrite (8) as follows:

$$P(V > v, X = L) \sim P(X = L)P(V^{s_L}_{M/M/1} > v)P\Big(L_{\text{res}} > \frac{v}{s_L(\mu s_L/\lambda - 1)}\Big),$$

as $v \to \infty$, where $V^{s_L}_{M/M/1}$ is the workload of the $M/M/1$ queue with arrival rate λ, service rate μ and one constant service speed s_L. This result has the following intuitive interpretation. It is well-known from standard large deviations theory that the most probable way for the workload $V^{s_L}_{M/M/1}$ in an $M/M/1$ queue with one constant speed s_L to get large is in a linear fashion, with a positive drift $s_L(\mu s_L/\lambda - 1)$ (see, e.g., p. 276 of Shwartz and Weiss (1995)). Hence the time until $V^{s_L}_{M/M/1}$ reaches the level v equals $v(s_L(\mu s_L/\lambda - 1))^{-1}$. So we need L_{past}, which has the same distribution as L_{res}, to have lasted at least $v(s_L(\mu s_L/\lambda - 1)^{-1}$. It is interesting to see that the rare event in (8) occurs as a result of two very different types of large deviations.

Finally consider (6). We can rewrite it as follows:

$$\begin{aligned} &P(V > v, X = H) \\ &\sim \quad P(X = H)P(V^{s_L}_{M/M/1} > v)P\Big(L > \frac{v}{s_L(\mu s_L/\lambda - 1)}\Big)\frac{\nu}{s_H - s_L}\frac{1}{\mu s_L - \lambda}, \end{aligned}$$

$v \to \infty$. The explanation is similar as above. Again the most probable way for the $M/M/1$ workload to grow is in a linear fashion, during a long low-speed period. The multiplicative factor $\nu(s_H - s_L)^{-1}(\mu s_L - \lambda)^{-1}$ represents the behaviour in subsequent high- and low-speed periods until the high-speed period under consideration. Notice that the occurrence of $s_H - s_L$ in the denominator makes sense: If s_L approaches s_H, then the system becomes an ordinary $M/M/1$ queue with a different workload tail behaviour. The occurrence of $\mu s_L - \lambda$ in the denominator also makes sense: If s_L approaches λ/μ, then the workload tail behaviour also changes.

3 Proof of Theorem 1

In both cases (i) and (ii) of Theorem 1, we shall derive the tail behaviour of the distribution functions $F_H(v)$, $F_L(v,\eta)$ and $F_L(v)$ from the asymptotic expansions of their LST in a neighbourhood of their first singularities: $\omega = 0$ in the case (i) and $\omega = \lambda/s_L - \mu$ in the case (ii). So, let us introduce the following LST:

$$\begin{aligned}\Phi_H(\omega) &:= \int\limits_{v=0-}^{\infty} e^{-\omega v} \mathrm{d}F_H(v), \qquad \Phi_L(\omega,\eta) := \int\limits_{v=0-}^{\infty} e^{-\omega v} \mathrm{d}F_L(v,\eta), \\ \Phi_L(\omega) &:= \int\limits_{v=0-}^{\infty} e^{-\omega v} \mathrm{d}F_L(v) = \int\limits_{\eta=0}^{\infty} \Phi_L(\omega,\eta)\mathrm{d}\eta.\end{aligned}$$

Starting point in our analysis is the following result, which was obtained in Boxma and Kurkova (1999) for the more general case of the $M/G/1$ queue with two service speeds: for all ω such that $\Phi_H(\omega) < \infty$ and

$$\delta(\frac{\omega}{\mu+\omega}(\lambda - s_L(\mu+\omega))\big) < \infty, \tag{12}$$

the LST $\Phi_H(\omega)$ satisfies the following equation

$$\Phi_H(\omega)k(\omega) = -s_H\omega F_H(0) - s_L\omega R(\omega), \tag{13}$$

where

$$k(\omega) \;:=\; \nu - \nu\delta(\,-f(\omega)) + \lambda\frac{\omega}{\mu+\omega} - s_H\omega,$$

$$\begin{aligned} R(\omega) &:= \int\limits_{\eta=0}^{\infty} e^{f(\omega)\eta} \int\limits_{x=0}^{\eta} e^{-f(\omega)x} \frac{F_L(0,x)}{1-L(x)} \mathrm{d}x \, \mathrm{d}L(\eta), \\ f(\omega) &:= \frac{\omega}{\mu+\omega}(s_L(\mu+\omega)-\lambda). \end{aligned}$$

Note that by Assumption 1, Equation (12) necessarily holds in $\{\omega : \mathrm{Re}\, f(\omega) < 0 \text{ or } |f(\omega)| < \varepsilon_0\}$. Moreover

$$\Phi_L(\omega,\eta) = (1-L(\eta))e^{f(\omega)\eta}\Big[\nu\Phi_H(\omega) - s_L\omega \int\limits_{x=0}^{\eta} e^{-f(\omega)x} \frac{F_L(0,x)}{1-L(x)} \, \mathrm{d}x\Big], \quad (14)$$

$$\Phi_L(\omega) = \nu\Phi_H(\omega)\frac{\delta(\,-f(\omega)) - 1}{f(\omega)} + \frac{s_L\omega(P(V=0, X=L) - R(\omega))}{f(\omega)} (15)$$

Formula (14) follows from (15) in Boxma and Kurkova (1999). Formula (15) is obtained from (14) via integration.

Case (i). Since $\lambda/\mu > s_L$, (12) holds for all $\omega > 0$. Under Assumption 1 one can find the asymptotics of $\Phi_H(\omega)$, $\Phi_L(\omega,\eta)$ and $\Phi_L(\omega)$ as $\omega \downarrow 0$ from Equation (13) proceeding along the lines of the proof of Theorem 4.1(i) in [8].

Then Theorem 8.1.6 of Bingham, Goldie and Teugels (1987) provides immediately the results (3), (4) and (5). The whole procedure is completely analogous to the proof of Theorem 4.1(i) in Boxma and Kurkova (1999), and therefore we omit its details.

Case (ii). For some $c > 0$ we have the following representation:

$$\frac{1}{1+\nu\delta} - F_H(v) = \int\limits_{c-i\infty}^{c+i\infty} e^{\omega v} \frac{\Phi_H(0) - \Phi_H(\omega)}{\omega} \, \mathrm{d}\omega. \qquad (16)$$

We shall show that the function $(\Phi_H(0) - \Phi_H(\omega))/\omega$ is analytic in $\{\omega : \mathrm{Re}\,\omega > \lambda/s_L - \mu - \delta\} \setminus \{\omega = \lambda/s_L - \mu\}$ for some $\delta > 0$ (Step 1). Then we shall find its expansion in the neighbourhood of its singularity $\omega = \lambda/s_L - \mu$ (Step 2). Finally, by virtue of Theorem 1 in Sutton (1934) applied to (16), we shall derive from this expansion the asymptotics (6) of the tail of $F_H(v)$.

Step 1. Analyticity. First of all, let us compare our two-speed system with the fixed-speed system $M/M/1$ having the same arrival and service characteristics but only one constant service speed s_L. Since $\lambda/\mu < s_L$, this fixed-speed system is stable. Let us denote its steady-state workload distribution by $V^{s_L}_{M/M/1}$. It is well-known that $P(V^{s_L}_{M/M/1} > v) = \lambda\mu^{-1}s_L^{-1}\exp\{(\lambda/s_L - \mu)v\}$. Moreover, for any given realisation of arrivals and services the workload in our two-speed system does not exceed the workload in this fixed-speed system at any moment of time. Then by simple arguments of stochastic ordering,

$$1 - F_H(v)/F_H(\infty) \le P(V^{s_L}_{M/M/1} > v) = \frac{\lambda}{\mu s_L}\exp\{(\lambda/s_L - \mu)v\}.$$

It follows that $(\Phi_H(0) - \Phi_H(\omega))/\omega$ is analytic at least in $\{\omega : \mathrm{Re}\,\omega > \lambda/s_L - \mu\}$ and by the same reasoning so are $(\Phi_L(0,\eta) - \Phi_L(\omega,\eta))/\omega$ and $(\Phi_L(0) - \Phi_L(\omega))/\omega$.

Let us denote the ε-neighbourhood of the point $\omega = \lambda/s_L - \mu$, excluding the point itself, by

$$A(\varepsilon) := \{\omega : |\omega - \lambda/s_L + \mu| < \varepsilon\} \setminus \{\omega = \lambda/s_L - \mu\}.$$

By the definition of $f(\omega)$, for all sufficiently small $\varepsilon > 0$,

$$A(\varepsilon) \subset \{\omega : |f(\omega)| < \varepsilon_0\} \setminus \{f(\omega) = 0\}, \tag{17}$$

where ε_0 satisfies (ii) and (iii) of Assumption 1. Moreover, since $k(\lambda/s_L - \mu) = (s_L - s_H)(\lambda/s_L - \mu) \neq 0$, then $k(\omega) \neq 0$ in $A(\varepsilon)$ for all sufficiently small $\varepsilon > 0$. Let us fix such a small $\varepsilon > 0$ ensuring (17) and prove that $\Phi_H(\omega)$ can be analytically continued to $A(\varepsilon)$. Due to (17) and Assumption 1, $\delta\{-f(\omega)\} < \infty$ for all $\omega \in A(\varepsilon)$. Then $\Phi_H(\omega)$ can be continued to $A(\varepsilon)$ by Equation (13):

$$\Phi_H(\omega) = \frac{-s_H\omega F_H(0) - s_L\omega R(\omega)}{k(\omega)} \qquad \text{for all } \omega \in A(\varepsilon). \tag{18}$$

Let us show that all functions in the right-hand side of (18) are analytic in $A(\varepsilon)$. The function $k(\omega) \neq 0$ is analytic in $A(\varepsilon)$ by its definition and (17). Namely,

$$\begin{aligned} k(\omega) \;\;&=\;\; -f(\omega)[\delta + (-f(\omega))^{\zeta-1}[C_L\Gamma(1-\zeta) + l(-f(\omega))] + h(-f(\omega))] \\ &\quad\; + (\lambda\omega)/(\mu+\omega) - s_H\omega, \end{aligned}$$

where $l(\omega)$ and $h(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon_0\}$. It is not difficult to prove that the function $R(\omega)$ is analytic in $A(\varepsilon)$ as well. For that purpose, let us note that for all ω

$$f(\omega) = f(\alpha(\omega)), \quad \text{where } \alpha(\omega) := \frac{\mu(\lambda - s_L(\mu+\omega))}{s_L(\mu+\omega)}. \tag{19}$$

Then $R(\omega) = R(\alpha(\omega))$ for all ω by the definition of $R(\omega)$. If $\omega \in A(\varepsilon)$, then $\alpha(\omega)$ lies in the neighbourhood of zero, where $\Phi_H(\omega)$ is well-defined and analytic. Hence, $\Phi_H(\alpha(\omega))$ is analytic in $A(\varepsilon)$. In addition, $|f(\omega)| = |f(\alpha(\omega))| < \varepsilon_0$ for all $\omega \in A(\varepsilon)$. Thus, if $\omega \in A(\varepsilon)$, Equation (13) is satisfied for $\alpha(\omega)$. It yields

$$\begin{aligned} s_L R(\omega) = s_L R(\alpha(\omega)) &= \frac{-\Phi_H(\alpha(\omega))k(\alpha(\omega))}{\alpha(\omega)} - s_H F_H(0) \\ = \;\; -\Phi_H(\alpha(\omega))\Big(\nu[\delta &+ (-f(\omega))^{\zeta-1}(C_L\Gamma(1-\zeta) + l(-f(\omega))) \\ &+ h(-f(\omega))]s_L\omega/\mu + s_L(\mu+\omega)/\mu - s_H\Big) - s_H F_H(0) \end{aligned} \tag{20}$$

for all $\omega \in A(\varepsilon)$. All functions in the right-hand side of (20) are analytic in $A(\varepsilon)$, whence so is $R(\omega)$. Consequently, since the right-hand side of Equation (18) is analytic in $A(\varepsilon)$, we conclude that $\Phi_H(\omega)$ is analytic there as well. Next, let us note that

$$\operatorname{Re} f(\lambda/s_L - \mu - \delta + iy) = \frac{(-\mu-\delta)s_L y^2 + s_L\delta(\mu+\delta-\lambda/s_L)(\lambda/s_L - \delta)}{(\lambda/s_L-\delta)^2 + y^2}. \tag{21}$$

Then for given $\varepsilon > 0$ there exists a $\delta > 0$ such that the domain $\{\omega : \lambda/s_L - \mu - \delta < \operatorname{Re}\omega < \lambda/s_L - \mu + \delta\} \setminus A(\varepsilon) \subset \{\omega : \operatorname{Re} f(\omega) < 0\}$. Indeed, $\delta(-f(\omega)) < \infty$ in this domain and again one can continue $\Phi_H(\omega)$ by Equation (13), where all the functions are analytic in the right-hand side. Thus, we have proved the analyticity of $\Phi_H(\omega)$ in the whole domain $\{\omega : \operatorname{Re}\omega > \lambda/s_L - \mu - \delta\} \setminus \{\omega = \lambda/s_L - \mu\}$. Let us also note that the equations (14) and (15), together with the analyticity of $\Phi_H(\omega)$ and $R(\omega)$, imply the analyticity of $\Phi_L(\omega, \eta)$ and $\Phi_L(\omega)$.

Recall, that our aim is to apply Theorem 1 of Sutton (1934) to the function $\psi(\omega) = (\Phi_H(0) - \Phi_H(\omega))/\omega$. We state this theorem in the Appendix for the reader's convenience. But its condition requires that the function $(\Phi_H(0) - \Phi_H(\omega))/\omega$, where $\omega = x + iy$, tends to zero as $y \to \pm\infty$ uniformly in $\lambda/s_L - \mu - \delta \le x \le \lambda/s_L - \mu + \delta$, for some $\delta > 0$ and in such a manner that the integral $\int_{-\infty}^{\infty} |(\Phi_H(0) - \Phi_H(x+iy))/(x+iy)|\,\mathrm{d}y$ converges at

$y = \pm\infty$. This assumption is not satisfied in our case. However, it is not difficult to see from the proof of the theorem proposed in Sutton (1934) that this assumption is not compulsory and can be relaxed, see the discussion in the appendix. Namely, one can replace it by the assumption that

$$|\int_{-\infty}^{\infty} e^{ity} \frac{(\Phi_H(0) - \Phi_H(x+iy))}{x+iy} \mathrm{d}y| \leq K$$

for some constant $K > 0$ and all sufficientlly large $t > 0$. But (using partial integration) for that it suffices to check only that

$$|\frac{(\Phi_H(0) - \Phi_H(x+iy))}{x+iy}| \leq \frac{K_1}{y} \qquad (22)$$

$$|\frac{d}{dy}\frac{(\Phi_H(0) - \Phi_H(x+iy))}{x+iy}| \leq \frac{K_2}{y^2} \qquad (23)$$

for some constants $K_1, K_2 > 0$, $\lambda/s_L - \mu - \delta \leq x \leq \lambda/s_L - \mu + \delta$ and all sufficiently large y. Using Equation (13) this a tedious but straightforward computation.

Let us also note that the equations (14) and (15), together with the analyticity of $\Phi_H(\omega)$ and $R(\omega)$, imply the analyticity of $\Phi_L(\omega, \eta)$ and $\Phi_L(\omega)$ in the domain $\{\mathrm{Re}\,\omega > \lambda/s_L - \mu - \delta\} \setminus \{\omega = \lambda/s_L - \mu\}$. The conditions analogous to (22) and (23) hold.

Step 2. Expansions. Now we derive the expansion of $\Phi_H(\omega)$ in $A(\varepsilon)$. This is equivalent to finding the expansion of $\Phi_H(\omega + \lambda/s_L - \mu)$ in $\{\omega : |\omega| < \varepsilon\}$. Equation (13) can be written as

$$\frac{\Phi_H(\omega + \lambda/s_L - \mu)}{\omega + \lambda/s_L - \mu} = \frac{-s_H F_H(0) - s_L R(\omega + \lambda/s_L - \mu)}{k(\omega + \lambda/s_L - \mu)}, \qquad (24)$$

for all ω with $|\omega| < \varepsilon$. By Assumption 1 and (17) we may write

$$\begin{aligned} & k(\omega + \lambda/s_L - \mu) \\ &= \nu\Big(1 - \delta\Big(-\frac{s_L\omega(\omega + \lambda/s_L - \mu)}{\lambda/s_L + \omega}\Big)\Big) + \lambda\frac{\omega + \lambda/s_L - \mu}{\lambda/s_L + \omega} - s_H(\omega + \lambda/s_L - \mu) \\ &= (\lambda/s_L - \mu)(s_L - s_H) + \omega h_1(\omega) + \omega^{\zeta} l_1(\omega), \qquad (25) \end{aligned}$$

where $h_1(\omega)$ and $l_1(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon\}$. To find the expansion of $R(\omega + \lambda/s_L - \mu)$, we note that

$$\alpha(\omega + \lambda/s_L - \mu) = -\frac{\mu\omega}{\lambda/s_L + \omega}.$$

Then, taking into account the analyticity of $\Phi_H(\omega)$ in the neighbourhood of $\omega = 0$, Assumption 1 and (17), we obtain:

$$\begin{aligned}
s_L R(\omega + \lambda/s_L - \mu) &= s_L R(\alpha(\omega + \lambda/s_L - \mu)) \\
&= -\frac{\Phi_H(\alpha(\omega + \lambda/s_L - \mu))k(\alpha(\omega + \lambda/s_L - \mu))}{\alpha(\omega + \lambda/s_L - \mu)} - s_H F_H(0) \\
&= \Big[\Phi_H\Big(-\frac{\mu\omega}{\lambda/s_L + \omega}\Big)\nu\Big(1 - \delta\Big\{-\frac{s_L\omega(\omega + \lambda/s_L - \mu)}{\lambda/s_L + \omega}\Big\}\Big) \\
&\qquad\qquad - \frac{\mu\omega(s_L(\lambda/s_L + \omega)/\mu - s_H)}{\lambda/s_L + \omega}\Big]\frac{\lambda/s_L + \omega}{\mu\omega} \\
&\quad - s_H F_H(0) \\
&= \Phi_H(0)[(s_H - \lambda/\mu) + \nu\delta(s_L - \lambda/\mu)] - s_H F_H(0) + \omega h_2(\omega) \qquad (26) \\
&\quad - \nu\Phi_H(0)(\lambda/\mu - s_L)C_L\Gamma(1-\zeta)\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}\omega^{\zeta-1}[1 + \omega l_2(\omega)]
\end{aligned}$$

for all ω with $|\omega| < \varepsilon$, where $l_2(\omega)$ and $h_2(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon\}$. Substitution of (25) and (26) into (24) gives

$$\begin{aligned}
\frac{\Phi_H(\omega + \lambda/s_L - \mu)}{\omega + \lambda/s_L - \mu} &= h_3(\omega) \qquad (27) \\
&\quad + \frac{\nu s_L C_L\Gamma(1-\zeta)}{(1+\nu\delta)(s_L - s_H)\mu}\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}\omega^{\zeta-1}[1 + \omega l_3(\omega)],
\end{aligned}$$

where $h_3(\omega)$ and $l_3(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon\}$. Therefore

$$\begin{aligned}
\omega\frac{1 - \Phi_H(\omega + \lambda/s_L - \mu)}{\omega + \lambda/s_L - \mu} &= \omega h_4(\omega) \qquad (28) \\
&\quad + \frac{\nu s_L}{(1+\nu\delta)(s_H - s_L)\mu}C_L\Gamma(1-\zeta)\Big(s_L(\mu s_L/\lambda - 1)\Big)^{\zeta-1}\omega^{\zeta}[1 + \omega l_4(\omega)],
\end{aligned}$$

where $h_4(\omega)$ and $l_4(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon\}$. Finally Theorem 1 of Sutton (1934) applies to the representation (16), where the integrand satisfies (28), and the proof of (6) is complete.

The result (7) can be derived analogously from equation (14) using the expansion (27) just obtained.

Let us proceed with $\Phi_L(\omega)$. It follows from Equation (15) that

$$\frac{\Phi_L(\omega+\lambda/s_L-\mu)}{\omega+\lambda/s_L-\mu} = I_1(\omega)+I_2(\omega), \tag{29}$$

where

$$\begin{aligned} I_1(\omega) &= \nu\frac{\Phi_H(\omega+\lambda/s_L-\mu)}{\omega+\lambda/s_L-\mu}\frac{\delta\{-f(\omega+\lambda/s_l-\mu)\}-1}{f(\omega+\lambda/s_L-\mu)}, \\ I_2(\omega) &= \frac{s_L P(V=0,X=L)-s_L R(\omega+\lambda/s_L-\mu)}{f(\omega+\lambda/s_L-\mu)}. \end{aligned}$$

Using the expansion (27) and also Assumption 1 and (17),

$$I_1(\omega) = h_5(\omega)+\omega^{\zeta-1}l_5(\omega), \tag{30}$$

for all ω with $|\omega|<\varepsilon$, where $h_5(\omega)$ and $l_5(\omega)$ are analytic in $\{\omega : |\omega|<\varepsilon\}$. The term $I_2(\omega)$ brings the main contribution to (29). Its expansion comes from (26), where the constant equals exactly $s_L P(V=0,X=L)$. In fact, by the definition of $R(\omega)$,

$$R(\lambda/s_L-\mu) = R(0) = \int_0^\infty F_L(0,x)\,\mathrm{d}x = P(V=0,X=L).$$

From the other point of view, substituting $\omega=0$ into (13), we get

$$s_L R(0) = \Phi_H(0)[(s_H-\lambda/\mu)+\nu\delta(s_L-\lambda/\mu)]-s_H F_H(0).$$

Since

$$f(\omega+\lambda/s_L-\mu) = \frac{s_L\omega(\omega+\lambda/s_L-\mu)}{(\lambda/s_L+\omega)},$$

then using (26)

$$\begin{aligned} I_2(\omega) = \Big[&\omega h_2(\omega) \\ &+\nu\Phi_H(0)(\lambda/\mu-s_L)C_L\Gamma(1-\zeta)\Big(s_L(\mu s_L/\lambda-1)\Big)^{\zeta-1}\omega^{\zeta-1}[1+\omega l_2(\omega)]\Big] \\ &\times\frac{(\lambda/s_L+\omega)}{s_L\omega(\omega+\lambda/s_L-\mu)} \\ = \quad &h_6(\omega)+\frac{\nu\lambda}{s_L\mu(1+\nu\delta)}C_L\Gamma(1-\zeta)\Big(s_L(\mu s_L/\lambda-1)\Big)^{\zeta-1}\omega^{\zeta-2}[1+\omega l_6(\omega)], \end{aligned}$$

where $h_6(\omega)$ and $l_6(\omega)$ are analytic in $\{\omega : |\omega| < \varepsilon\}$. Thus

$$\omega \frac{1-\Phi_L(\omega+\lambda/s_L-\mu)}{\omega+\lambda/s_L-\mu} = \omega h_7(\omega) \\ - \frac{\nu\lambda}{s_L\mu(1+\nu\delta)} C_L \Gamma(1-\zeta)\Big(s_L(\mu s_L/\lambda-1)\Big)^{\zeta-1}\omega^{\zeta-1}[1+\omega l_6(\omega)],$$

where $h_7(\omega)$ is analytic in $\{\omega : |\omega| < \varepsilon\}$. Finally Theorem 1 from Sutton (1934) together with this expansion implies the result (8).

Appendix

Theorem 2 *Let us define*

$$f(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{t\omega}\psi(\omega)\,\mathrm{d}\omega,$$

where $\omega = x+iy$ and the path of integration is the straight line $x = c$, chosen such that $\psi(\omega)$ is analytic for $x \geq c$. Let

1. *$\psi(\omega)$ be analytic for $\mathrm{Re}\,\omega \geq a-\delta$ for some $\delta > 0$ except at k points $\omega_1, \ldots, \omega_p, \ldots, \omega_k$ on $\mathrm{Re}\,\omega = a$;*
2. *near each such point ω_p we have*

$$(\omega-\omega_p)\psi(\omega) = \sum_{n=0}^{\infty} a_{np}(\omega-\omega_p)^n + (\omega-\omega_p)^{\beta_p}\sum_{n=0}^{\infty}\beta_{np}(\omega-\omega_p)^n,$$

where $0 < \beta_p < 1$, and the series converge for $|\omega-\omega_p| < l$, $l > 0$;

3. *$\psi(\omega) \to 0$ as $y \to \pm\infty$ uniformly in x for $a-\delta \leq x \leq c$, $(c > a)$ and in such a manner that $\int|\psi(\omega)|\mathrm{d}y$ converges at $y = \pm\infty$.*

Then for $t > 0$:

$$f(t) \sim \sum_{p=1}^{k} e^{\omega_p t}\Big(a_{0p} + \frac{\sin\pi\beta_p}{\pi}\sum_{n=0}^{\infty}(-1)^n b_{np}\Gamma(\beta_p+n)t^{-\beta_p-n}\Big). \qquad (31)$$

Proof and discussion. This theorem is proved by Sutton (1934). In his proof the integration path is deformed into a path from $a-\delta-i\infty$ to $a-\delta+i\infty$ composed of the line $x=a-\delta$ interrupted at the k points $(a-\delta, y_p)$, the k lines $y=y_p$ (each taken twice) from $x=a-\delta$ to $x=a-\varepsilon/t$ and the k circles $|\omega-\omega_k|=\varepsilon/t$. Thus by Cauchy's Theorem,

$$f(t)=\frac{1}{2\pi i}\int\limits_{a-\delta-i\infty}^{a-\delta+i\infty} e^{(a-\delta+iy)t}\psi(a-\delta+iy)\,\mathrm{d}y+\sum_{p=1}^{k}(I_p'+I_p''),$$

where I_p' and I_p'' are contributions from the double line $y=y_p$ and the circle with the center ω_p respectively. Sutton shows that by (ii) the asymptotics of the sum $\sum_{p=1}^{k}(I_p'+I_p'')$ is given by the right-hand side of (31), while the contribution of the integral along the straight line $\omega=a-\delta+iy$ is exponentially small:

$$\Big|\int\limits_{a-\delta-i\infty}^{a-\delta+i\infty} e^{(a-\delta+iy)t}\psi(a-\delta+iy)\,\mathrm{d}y\Big|\leq Ke^{at}e^{-\delta t}, \tag{32}$$

where $K>0$ is a constant. Condition (iii) of course suffices for (32) but it is not necessary. One can relax it and assume instead that

$$\Big|\int\limits_{a-\delta-i\infty}^{a-\delta+i\infty} e^{iyt}\psi(a-\delta+iy)\,\mathrm{d}y\Big|\leq K$$

for some constant $K>0$ and all sufficiently large $t>0$.

Acknowledgement. The authors are indebted to Ivo Adan, Sem Borst and Bert Zwart for interesting discussions.

References

[1] R.P. Agrawal, A.M. Makowski and Ph. Nain (1999). On a reduced-load equivalence for fluid queues under subexponentiality. *Queueing Systems* **33**, 5–41.

[2] N.H. Bingham, C.M. Goldie and J.L. Teugels (1987). *Regular Variation.* Cambridge Univ. Press, Cambridge.

[3] S.C. Borst, O.J. Boxma and P.R. Jelenković (1999). Generalized processor sharing with long-tailed traffic sources. In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key and D. Smith (North-Holland, Amsterdam), pp. 345-354.

[4] S.C. Borst, O.J. Boxma and P.R. Jelenković (1999). Coupled processors with regularly varying service times. Technical Report COSOR 99-11, Eindhoven University of Technology. To appear in: Proceedings of Infocom 2000.

[5] S.C. Borst and A.P. Zwart (2000). A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows. SPOR Report 2000-04 Eindhoven University of Technology.

[6] O.J. Boxma and J.W. Cohen (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems* **33**, 177-204.

[7] O.J. Boxma, Q. Deng and A.P. Zwart (1999). Waiting-time asymptotics for the $M/G/2$ queue with heterogeneous servers. Technical Report COSOR 99-20, Eindhoven University of Technology.

[8] O.J. Boxma and I.A. Kurkova (1999). The $M/G/1$ queue with two speeds. EURANDOM report 99-057, submitted to *Appl. Probab. Journals.*

[9] J.W. Cohen (1973). Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab.* **10**, 343–353.

[10] J.W. Cohen (1982). *The Single Server Queue.* North-Holland Publ. Co., Amsterdam.

[11] S. Halfin (1972). Steady-state distribution for the buffer content of an $M/G/1$ queue with varying service rate. *SIAM J. Appl. Math.* **23**, 356-363.

[12] D.P. Heyman and T.V. Lakshman (1996). Source models for VBR broadcast-video traffic. *IEEE/ACM Trans. Netw.* **4**, 40-48.

[13] P.R. Jelenković, A.A. Lazar and N. Semret (1997). The effect of multiple time scales and subexponentiality of MPEG video streams on queueing behavior. *IEEE J. Sel. Areas Commun.* **15**, 1052-1071.

[14] P.R. Jelenkovićand A.A. Lazar (1999). Asymptotic results for multiplexing subexponential on-off processes. *Adv. Appl. Probab.* **31**, 394–421.

[15] O. Kella and W. Whitt (1992). A storage model with a two-state random environment. *Operations Research* **40**, S257–S262.

[16] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson (1993). On the self-similar nature of Ethernet traffic. In: *Proc. SIGCOMM '93*, pp. 183-193.

[17] W. Li, D. Shi and X. Chao (1997). Reliability analysis of $M/G/1$ queueing systems with server breakdowns and vacations. *J. Appl. Probab.* **34**, 546-555.

[18] M.F. Neuts (1971). A queue subject to extraneous phase changes. *Adv. Appl. Probab.* **3**, 78-119.

[19] R. Núñez Queija (1998). Sojourn times in a processor-sharing queue with service interruptions. CWI Report PNA-R9807, Amsterdam; accepted for publication in *Queueing Systems*.

[20] K. Park and W. Willinger, eds. (2000). *Self-Similar Network Traffic and Performance Evaluation*. Wiley, New York.

[21] A. Shwartz and A. Weiss (1995). *Large Deviations for Performance Analysis*. Chapman and Hall, London.

[22] W.G.L. Sutton (1934). The asymptotic expansion of a function whose operational equivalent is known. *Journal of the London Mathematical Society* **9**, 131–137.

[23] T. Takine and B. Sengupta (1997). A single server queue with service interruptions. *Queueing Systems* **26**, 285-300.

[24] U. Yechiali and P. Naor (1971). Queueing problems with heterogeneous arrival and service. *Operations Research* **19**, 722-734.

[25] U. Yechiali (1973). A queueing-type birth-and-death process defined on a continuous-time Markov chain. *Operations Research* **21**, 604-609.